AI Researchers
Embedding Responsible AI in Research

AI READINESS TOOLKIT

UC3-03 Use case 3: Data preparation and content analysis

AI for hypothesis testing

A bounded methods-referee workflow for checking that a test matches the hypothesis

AT A GLANCE

A supervisor uses a bounded AI methods referee to check whether a proposed statistical test fits the hypothesis and the study design. The case gives the prompt, shows how to question the AI's reasoning and how to record a transparent decision that a person approves. It helps supervisors and researchers discuss test choice with confidence. Interpretation and conclusions stay with the researchers.

TYPE
Case study

FORMAT
PDF

CONTRIBUTING PARTNER
FEP, University of Porto

WHO IT IS FOR
Supervisors and researchers choosing and defending a statistical test

FRAMEWORK BARRIERS
Trust in AI, Know-how

HUMAN-ONLY BOUNDARY

No unpublished or sensitive data in unapproved tools. Interpretation and conclusions stay human.

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

What this use case is

This is a ready-to-use case study and guided demonstration for academic supervisors, research mentors and early-career researchers. It focuses on one durable research task: checking the fit between an investigation hypothesis and the proposed statistical test or model. The named AI tool may change; the reasoning sequence should remain stable.

The AI acts as a critical second reader of an analysis plan. It asks for missing design information, proposes a short list of methods, rejects plausible near misses and drafts a decision record. The supervisor or researcher verifies every claim and owns the final analysis decision.

User scenario

A doctoral researcher gives a supervisor a short analysis plan before preregistration or manuscript submission. The plan contains a clear scientific hypothesis but names a statistical test without explaining why it fits the design. Together they use an institution-approved AI environment to run a structured methods-referee check on the doctoral researcher's plan (or on a fictional plan).

> Why this belongs under refereeing? The activity reproduces a bounded methodological consistency check that a referee, supervisor or internal reviewer might perform. It does not ask AI to evaluate scientific merit, make an editorial recommendation or review a confidential third-party manuscript.

Intended outcomes

Translate a verbal hypothesis into a population, outcome, contrast and estimand before naming a test.

Distinguish independent, paired, repeated and clustered observations.

Use outcome scale, design, covariates, sample structure and assumptions to narrow a method shortlist.

Ask AI to expose missing information and reject plausible but unsuitable tests.

Document an effect measure, uncertainty, diagnostics, limitations and the final human decision.

Responsible-use boundary

BOUNDARY	RULE
AI may support	Ask clarifying questions; identify inconsistencies; offer alternatives; draft a checklist or decision record; explain why a candidate method may fail.
Human must decide	Scientific question; estimand; causal or associational interpretation; admissible assumptions; final method; data exclusions; sensitivity analyses; claims and conclusions.
Do not provide	Unpublished third-party manuscripts, confidential referee material, identifiable data, restricted data or proprietary content to an external AI service.
Escalate	Complex surveys, multilevel or longitudinal designs, adaptive trials, small samples, causal inference, extensive missingness, multiplicity, regulated or high-stakes analyses.

Use the fictional case, a synthetic abstraction, or the researcher's own authorised analysis plan in an institution-approved environment.

Record an initial human choice before consulting AI, so automation bias can be discussed rather than hidden.

Preserve unknowns. AI must not invent design details, data distributions, references, power calculations or assumption checks.

Do not select a test from the hypothesis sentence alone. Supply the sampling/allocation, unit of analysis, outcome scale, timing, pairing/clustering, covariates and missing-data context.

Do not report only a test name and p-value. Require the effect estimate, uncertainty interval, diagnostics and limitations.

The methods-referee workflow

STEP	ACTION
1. Specify	Rewrite the question as population + outcome + contrast + estimand.
2. Commit	Record the researcher's pre-AI method choice and rationale.
3. Elicit	Ask AI for questions only; mark each answer supplied, unknown or not applicable.
4. Shortlist	Request one primary method, credible alternatives and rejected near misses.
5. Red-team	Challenge dependence, assumptions, missingness, multiplicity and interpretation.
6. Decide	Complete the decision record; accept, revise, reject or escalate each material suggestion.

Fictional case: an AI-supported statistics clinic

A university evaluates an AI-supported statistics clinic for doctoral researchers. Eighty volunteers are individually randomised before the course to the AI-supported clinic or standard self-study. Statistical knowledge is measured on a 0–100 scale immediately before allocation and again after four weeks.

> Primary investigation hypothesis: Participants offered the AI-supported clinic will have higher post-course statistical knowledge than participants offered standard self-study, after accounting for baseline knowledge.

FIELD	STUDY INFORMATION
Population	Doctoral researchers eligible for the four-week course
Allocation	Individual 1:1 randomisation; no seminar- or site-level allocation
Primary outcome	Post-course knowledge score, continuous 0–100
Baseline measure	Pre-course knowledge score, continuous 0–100; pre-specified adjustment
Primary contrast	Offer of AI-supported clinic versus standard self-study
Target quantity	Adjusted mean difference in post-course knowledge between allocated groups

FIELD	STUDY INFORMATION
Direction	Two-sided primary analysis unless a directional test was justified before data inspection
Known limitation	Some post-course scores may be missing; reasons and group-specific missingness are not yet known

THE SUBMITTED ANALYSIS-PLAN EXCERPT

> Text for review: “The hypothesis will be tested with a paired-samples t-test comparing pre- and post-course knowledge in the AI-supported group. Statistical significance at $p < .05$ will confirm that the clinic caused an improvement.”

COMMIT BEFORE CONSULTING AI

Individually write a provisional answer. The aim is to make any later change in judgement visible.

> Referee's first question: Does the proposed paired comparison estimate the between-group effect stated in the primary hypothesis — or only change within one selected group?

Guided AI demonstration

Run the prompts in sequence. Do not start by asking “Which test should I use?” Each prompt constrains the AI to a review role and creates an auditable intermediate output.

PROMPT 1 — CLARIFY BEFORE RECOMMENDING

> Act as a statistical methods referee, not as the final decision-maker. Read the fictional hypothesis and analysis-plan excerpt below. Do not recommend a test yet. Ask only the questions needed to determine: 1) the population, outcome, contrast and estimand; 2) allocation/sampling and unit of analysis; 3) whether observations are independent, paired, repeated or clustered; 4) outcome scale, timing, covariates and missing-data context; 5) whether the planned analysis actually tests the stated hypothesis. For every unresolved point, write UNKNOWN rather than assuming. [Paste only the fictional case in Section 6.]

PROMPT 2 — PRODUCE AND CRITIQUE A SHORTLIST

> Using only the confirmed information, prepare a methods-referee note with: A. the research question rewritten as population + outcome + contrast + estimand; B. one primary method and why it targets that estimand; C. up to two credible alternatives and the conditions under which they would be preferable; D. three plausible near-miss methods that should be rejected, with reasons; E. assumptions/diagnostics, missing-data concerns and what must be reported; F. an explicit uncertainty statement and any reason to consult a statistician. Do not infer data distributions or claim that assumptions have been checked.

PROMPT 3 — RED-TEAM THE RECOMMENDATION

> #### NOW ACT AS A SCEPTICAL SECOND REFEREE
> Try to falsify the proposed primary method. Check: estimand mismatch; wrong unit of analysis; ignored dependence; post-randomisation adjustment; model-form assumptions; influential observations; missing outcomes; multiplicity; and causal overstatement. Return a table: issue | why it matters | evidence needed | keep/revise/reject/escalate. End with the sentence: “The final methodological decision and interpretation remain human responsibilities.”

Required output: Statistical Test Decision Record

The activity is complete only when a human reviewer can reconstruct why the proposed method was accepted, revised, rejected or escalated. Copy the fields below into the final record.

RECORD FIELD	REQUIRED CONTENT
Research question	Population, outcome, comparison/association and time horizon
Primary estimand	Exact target quantity and effect scale
Design	Sampling/allocation, unit of analysis, pairing/repetition/clustering
Outcome	Scale, support, timing and relevant distributional features
Primary method	Test/model, model formula in words, contrast and justification
Rejected near misses	Method + reason it does not answer the stated question
Assumptions/diagnostics	What must be checked; what will be done if concerns arise
Missing data	Extent/reasons to inspect; primary handling and sensitivity analysis
Multiplicity	Primary versus secondary outcomes; any adjustment or hierarchy
Reporting	Effect estimate, confidence interval, exact sample accounting and p-value where relevant
AI audit	Accepted, revised, rejected and escalated AI suggestions
Human approval	Reviewer name/role, date, decision and unresolved limitations

AI DECISION LOG

AI SUGGESTION	DECISION	REASON / EVIDENCE	OWNER
	Accept / Revise / Reject / Escalate		
	Accept / Revise / Reject / Escalate		
	Accept / Revise / Reject / Escalate		
	Accept / Revise / Reject / Escalate		

> Minimum reporting standard: Name the estimand and effect measure, report an uncertainty interval, account for analysed and missing observations, document diagnostics and distinguish statistical compatibility from scientific importance. A p-value alone is not the result.

Facilitator benchmark

> Defensible primary choice: An intention-to-treat linear regression / ANCOVA of post-course knowledge on allocated group and pre-course knowledge, reporting the adjusted between-group mean difference with a confidence interval.

ELEMENT	BENCHMARK REASONING
Why it fits	The stated estimand is a between-group difference at follow-up, adjusted for a pre-randomisation prognostic measure. Individual randomisation supports individual-level analysis.
Model in words	Post-course score = intercept + allocated group + baseline score; optionally include pre-specified randomisation strata.
Primary effect	Coefficient for allocated group: adjusted mean difference in post-course knowledge.
Core checks	Residual patterns, linearity of baseline relationship, influential observations, treatment-by-baseline interaction if scientifically relevant, and robustness of standard errors.
Missing outcomes	Describe extent and reasons by allocated group; justify primary handling; add an appropriate sensitivity analysis rather than assuming missingness is harmless.
Interpretation	Randomisation supports causal interpretation of the offer under the study design and assumptions; the result does not establish that AI caused learning for every participant.

PLAUSIBLE NEAR MISSES

CANDIDATE	REFEREE JUDGEMENT
Paired t-test within the AI group	Tests change inside one group; it does not estimate the between-group effect and cannot separate intervention change from background change.
Post-course independent t-test only	Targets a between-group difference, but discards a pre-specified prognostic baseline measure and may be less precise. It may be a simple secondary check, not the best primary plan here.
t-test of individual change scores	Can answer a between-group change contrast, but targets a different formulation and is often less efficient than baseline-adjusted follow-up when baseline predicts outcome.
Repeated-measures ANOVA	Can represent group-by-time change, but adds structure not needed for a single follow-up estimand and may obscure the simpler adjusted comparison.
Two separate within-group tests	“Significant in one group, not significant in the other” is not evidence of a significant group difference.

> Important nuance: The benchmark is not a universal recipe. A different estimand, non-linear outcome, cluster allocation, repeated follow-ups, small-sample problem or severe model failure can change the appropriate method.

Transfer branches: change one design feature

For each branch, revise the estimand first, then predict how the method changes. Use the same three-prompt workflow only after writing a human answer.

BRANCH	CHANGED CONDITION	EXPECTED DIRECTION
A. One group, two occasions	All participants receive the clinic; compare pre/post continuous scores.	Paired analysis or regression for within-person change; no randomised control-group effect.
B. Binary outcome	Primary outcome becomes course completion (Yes/No).	Binomial model or appropriate two-group proportion analysis; report a pre-specified risk-based effect where possible.
C. Three follow-ups	Knowledge is measured at baseline, week 4, week 8 and week 12.	Longitudinal model that accounts for within-person correlation and represents time appropriately.
D. Cluster allocation	Seminar groups — not individuals — are randomised.	Cluster-aware analysis, such as a multilevel model or cluster-level method; unit-of-analysis error becomes central.
E. Ordinal paired outcome	Confidence is rated 1–5 before and after the course.	An ordinal model or a clearly justified paired ordinal procedure; do not automatically treat five categories as continuous.

Evaluation rubric

CRITERION	POINTS	FULL-CREDIT EVIDENCE
Question and estimand	4	Population, outcome, contrast and target quantity are explicit and aligned.
Design and dependence	4	Unit of analysis, allocation, pairing/repetition/clustering and covariates are correctly identified.
Method reasoning	4	Primary method is justified; alternatives and near misses are accurately distinguished.
Verification plan	4	Assumptions, diagnostics, missingness, multiplicity and escalation are addressed.
Reporting and boundary	4	Effect + interval + sample accounting are planned; human approval and AI audit are explicit.

> **SUGGESTED INTERPRETATION 17–20**
> defensible and auditable; 13–16: revise before approval; 9–12: substantial methodological gaps; 0–8: do not proceed without expert support.

DEBRIEF QUESTIONS

> What did the AI ask that the original plan omitted?

Did the AI's confidence exceed the available evidence?

Which rejected method looked most plausible, and why was it wrong for the stated estimand?

What changed between the initial human choice and final decision?

Where would institutional rules or confidentiality prevent use of an external AI service?

Risks and mitigations

RISK	MITIGATION BUILT INTO THIS USE CASE
Automation bias	Record the human choice first; require rejected alternatives and an AI decision log.
Confidentiality breach	Use fictional/synthetic material or an approved environment; never paste confidential referee manuscripts into external tools.
Hallucinated facts or references	Preserve UNKNOWN; verify claims against the protocol, data dictionary and authoritative statistical sources.
Test-name matching	Force the sequence estimand → design → dependence → outcome → assumptions → reporting.
False certainty	Require uncertainty language, evidence needed and escalation triggers.
p-value fixation	Make effect estimate, confidence interval, sample accounting and substantive interpretation mandatory.

Suggested delivery

TIME	ACTIVITY
0–8 min	Read the case and record an initial human decision.
8–18 min	Specify question and estimand; identify the mismatch in the submitted plan.
18–35 min	Run the three prompts and complete the question log.
35–48 min	Compare with the benchmark; complete the decision record.
48–60 min	Apply one transfer branch and debrief the human–AI boundary.

The resource can be used as an individual case, a peer-led discussion, or a short AI demonstration.

About this resource

TITLE	AI for hypothesis testing
TOOLKIT REFERENCE	UC3-03, use case 3: Data preparation and content analysis
DEVELOPED BY	FEP University of Porto, Fábio Duarte
SUGGESTED CITATION	AI Researchers (2026). AI for hypothesis testing. AI Readiness Toolkit, UC3-03. FEP University of Porto, Fábio Duarte. airesearchers.eu
QUALITY ASSURANCE	Reviewed by the AI Researchers consortium for quality, categorisation and alignment with the AI Engagement Framework. External peer review by senior academics is under way, and the item may be updated as a result.
VERSION	1.0, November 2026
PROJECT	AI Researchers: Embedding Responsible AI in Research. Erasmus+ Cooperation Partnership in Higher Education, KA220-HED-66835D2B. October 2025 to March 2028.
CONSORTIUM	University of Lodz (coordinator), FEP University of Porto, UPF Barcelona School of Management, ACEEU, NIPI (OPI PIB), European E-learning Institute and Momentum Marketing Services

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

AI Researchers
Embedding Responsible AI in Research